Copy-number-aware differential analysis of quantitative DNA sequencing data

Robinson, Mark D ; Strbenac, Dario ; Stirzaker, Clare ; Statham, Aaron L ; Song, Jenny ; Speed, Terence P ; Clark, Susan J

Abstract: Developments in microarray and high throughput sequencing (HTS) technologies have resulted in a rapid expansion of research into epigenomic changes that occur in normal development and in the progression of disease, such as cancer. Not surprisingly, copy number variation (CNV) has a direct effect on HTS read densities and can therefore bias differential detection results. We have developed a flexible approach called ABCD-DNA (Affinity Based Copy-number-aware Differential quantitative DNA sequencing analyses) that integrates CNV and other systematic factors directly into the differential enrichment engine.

DOI: https://doi.org/10.1101/gr.139055.112

Copy-number-aware differential analysis of quantitative DNA sequencing data

Mark D Robinson, Dario Strbenac, Clare Stirzaker, et al.

Genome Res. published online August 9, 2012
Access the most recent version at doi:10.1101/gr.139055.112

Copy-number-aware differential analysis of quantitative DNA sequencing data

Mark D Robinson[1,2,3*], Dario Strbenac[3], Clare Stirzaker[3,6], Aaron L. Statham[3], Jenny Song[3], Terence P. Speed[4,5], Susan J. Clark[3,6]

[1]Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190 CH-8057 Zurich, Switzerland

[2]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland

[3]Epigenetics Laboratory, Cancer Research Program, Garvan Institute of Medical Research, Sydney 2010, New South Wales, Australia

[4]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Melbourne 3052, Victoria, Australia

[5]Department of Medical Biology, University of Melbourne, 3050, Victoria, Australia

[6]St Vincent's Clinical School, University of NSW, Sydney, NSW, Australia

[*]Corresponding author

Email addresses:

MDR: mark.robinson@imls.uzh.ch

DS: d.strbenac@garvan.org.au

CS: c.stirzaker@garvan.org.au

ALS: a.statham@garvan.org.au

JS: j.song@garvan.org.au

TPS: terry@wehi.edu.au

SJC: s.clark@garvan.org.au

Running title: Copy-number-aware differential analyses

Keywords: Statistics; Epigenomics; Next generation sequencing

ABSTRACT

Developments in microarray and high throughput sequencing (HTS) technologies have resulted in a rapid expansion of research into epigenomic changes that occur in normal development and in the progression of disease, such as cancer. Not surprisingly, copy number variation (CNV) has a direct effect on HTS read densities and can therefore bias differential detection results. We have developed a flexible approach called ABCD-DNA (**A**ffinity **B**ased **C**opy-number-aware **D**ifferential quantitative **DNA** sequencing analyses) that integrates CNV and other systematic factors *directly* into the differential enrichment engine.

INTRODUCTION

All normal cells carry the same DNA sequence, yet distinct cell types result from gene expression patterns that are controlled by a combination of genetic and epigenetic mechanisms. In cancer, genetic and epigenetic changes result in altered gene expression patterns, such as up-regulation of oncogenes and down-regulation of tumour-suppressor genes (Stratton 2011; Jones and Baylin 2007). Specifically, mutations in the DNA sequence or changes in copy number can alter how these genes are regulated or expressed, as can non-sequence epigenetic features, such as chemical (e.g. DNA methylation or histone modifications) or structural makeup (e.g. nucleosome occupancy). Advances in microarray and especially HTS technologies have driven a deeper exploration of genetic and epigenetic phenomena, resulting in several large data collection projects (Stratton 2011; Jones et al. 2008; Bernstein et al. 2010; International Cancer Genome Consortium 2010) as well as many smaller scale studies. Statistical and computational tools for processing and interpreting these datasets are maturing, and altogether these give exciting prospects for the understanding, detection, prevention, and treatment of cancer and other diseases.

Recently, we highlighted that comparisons between cancer and normal epigenomes need to be informed by genomic changes (Robinson, Statham, et al. 2010; Robinson, Clare Stirzaker, et al. 2010). Specifically, CNV has a direct effect on read densities of affinity- (or enrichment-) based assays (e.g. Chromatin immunoprecipitation, ChIP and methylated DNA capture, MBDCap); we refer to these techniques collectively as qDNA-seq, since they all provide a quantitative epigenetic readout at a specific loci. In these assays, a subset of target DNA fragments are captured, prepared, sequenced and mapped to a reference genome. Enrichment levels are interpreted as the relative abundance across two populations having the property of interest. Consider comparing enrichment levels between two prostate cell lines -- normal epithelial cells (PrEC) and cancer cells (LNCaP). There is significant CNV between PrEC and LNCaP cells, as shown in **Figure 1A** (see also **Supplementary Figure 1**). The CNV imbalance leads directly to changes in read density that are *not* reflective of true changes in methylation (e.g. from MBDCap-seq data). Using Illumina HumanMethylation 450k arrays as an independent assessment of changes in DNA methylation that should be unaffected by CNV (Houseman et al. 2009), **Figures 1B-E** highlight both false positive and false negative detections using existing algorithms; these examples are accurately detected by our ABCD-DNA approach (details below). Interestingly, because the prominent copy number state of LNCaP cells is 4 (**Figure 1a, Supplementary Figure 1**), depth-adjusted read densities are approximately "neutral" (in terms of sampling captured DNA) when LNCaP and PrEC cells have 4 and 2 copies, respectively; this further imbalance can be adjusted through "normalization" (adjustments for depth and diversity) in the statistical modeling.

There are now a large number of tools for *absolute* analysis of qDNA-seq data; methods are available for the detection of short distinct events (e.g. MACS (Yong

Zhang et al. 2008)), enriched *regions* (e.g. RSEG (Qiang Song and Smith 2011), ChromaBlocks (Hawkins et al. 2010), or both simultaneously with ZINBA (Rashid et al. 2011)). However, none of the tools are designed explicitly for *differential* analyses or for when replication is available. Recently, a framework called DiffBind was developed to post-process output from absolute algorithms into merged regions and perform differential analysis based on read densities (Ross-Innes et al. 2012).

A separate class of methods are available to *directly* detect differential regions, often without the use of input or other control samples (see Table 1 for list of assays and acronyms). For example, Bock *et al*. detected changes in read density using Fisher's exact test; CNV is deemed unimportant in their analysis despite no CNV-typing (Bock et al. 2010). Another strategy, ChIPDiff, assumes beta-binomially distributed tiled bin counts and uses a Hidden Markov Model (HMM) to combine adjacent differential regions (Xu et al. 2008). Similarly, RSEG scans for differential regions using an HMM with a difference-of-negative-binomials emission distribution (Qiang Song and Smith 2011). Other tools are emerging for differential analyses, such as DBChIP (Liang and Keles 2011) or by collecting existing Unix-based tools (Bardet et al. 2011), but none of these are explicitly CNV-aware. Though specific to DNA methylation, BATMAN, which transforms read densities into absolute methylation estimates, was recently made CNV-aware by first dividing read densities by copy number before differential analysis (Feber et al. 2011; Down et al. 2008). However, this transformation takes measurements off the count scale, which may affect the sensitivity of subsequent statistical analyses.

We propose a flexible and general statistical framework called ABCD-DNA that explicitly adjusts for CNV in differential epigenome analyses. First, we describe the statistical framework, which necessarily involves considerations for the estimation of

CNV and normalization. Second, we illustrate the effects of CNV on various algorithms for differential analysis across multiple qDNA-seq datasets. Using independent truth (DNA methylation levels), we demonstrate improved differential detection performance using CNV-aware analyses. Third, we compare the performance of ABCD-DNA and competing methods, demonstrating the proposed framework is competitive against existing approaches and flexible, irrespective of CNV compensation. All methods are freely available in public software projects and R scripts to reproduce all analyses are provided.

RESULTS

A general framework for CNV-aware differential qDNA-seq analyses

We propose the following framework:

1. Generate read counts at regions of interest (e.g. at detected peaks, tiled regions genome-wide, or proximal to transcription starts);
2. Estimate copy number offsets from an external data source (see "Copy number analyses" below);
3. Estimate normalization offsets based on CNV-neutral loci (See "Normalization" below);
4. Perform differential analysis of count data (e.g. using edgeR) using offsets.

Formally, the strategy for CNV-aware differential analyses can be encapsulated in a generalized linear model (GLM), where tools applicable to genome-scale datasets have recently become available (McCarthy et al. 2012; Anders and Huber 2010; Zhou et al. 2011). Specifically, let Y_{ij} be the read count for region of interest i in sample j (i=1,…,r and j=1,…,n where r is the number of regions and n is the number of samples). The read density observed at any genomic region is modified by systematic effects, such as "effective" sequencing depth, copy number, and underlying biological

factors of interest, as well as sampling and biological variability. Offsets impose a higher or lower expected mean based on the systematic factors, such as copy number state, depth of sequencing and sampling rates due to the diversity of the library sequenced; these are estimated in advance and treated as fixed in the downstream analysis. We model the logarithm of expected value of Y_{ij} as follows:

$$\log(E[Y_{ij}]) = O_{ij} + B_i X$$

where O_{ij} is an rxn matrix of offsets that match the count matrix, X is an rxk matrix that captures the experimental design (conditions, covariates) and B_i is a rxk matrix of region-specific coefficients. O_{ij} can be decomposed into $\log(CN_{ij}) + \log(1\ D_j)$ where CN_{ij} is a matrix of offsets for copy number and D_j represents sample-specific offset vector, both of which can be calculated as suggested above. To make inferences regarding differential enrichment, hypothesis tests can be formulated (e.g. likelihood ratio test) on the parameters of interest within the B_i matrix (e.g. cancer versus normal); tools for this are readily available (e.g. edgeR (Robinson, McCarthy, et al. 2010)). See Supplementary PDF Document for specification of all the modeling details (e.g. distributional assumptions, statistical testing).

ABCD-DNA can use alternative CNV sources; CNV linearly affects qDNA-seq

ABCD-DNA requires pre-processed CNV information to be delivered to a GLM in a corresponding matrix for regions of interest for each sample; in theory, our approach is independent of the source of CNV information. However, in practice, the success of the CNV adjustment will be determined by the accuracy, resolution and scale of the CNV estimates, which can vary widely with the platform and preprocessing algorithm used (Curtis et al. 2009). Accuracy should be facilitated by smoothing techniques, such as segmentation (Venkatraman and Olshen 2007) while resolution is ultimately

determined by probe spacing (microarrays) or depth of sequencing (HTS). In our analysis of PrEC and LNCaP cells, we used the PICNIC algorithm on Affymetrix SNP 6.0 array data, which resulted in integer-valued CNV estimates due to the homogenous population of the cell lines (**Figure 1A**). **Supplementary Figure 2** highlights strong concordance between PICNIC CNV estimates and segmented low-coverage genomic sequencing read densities, after adjusting for GC content and mappability (See Methods). Therefore, only minor differences in the downstream differential analysis between the alternative sources of CNV offsets should result (discussed below). Another important consideration is the scale of the CNV offsets, and specifically, the relationship between CNV and DNA-seq read depths; the GLM model assumes a linear relationship between the offset and expected mean. **Supplementary Figure 3** shows M (log-fold-change adjusting for total depth) versus A (average-log-read-density) "smear" plots for three qDNA-seq datasets across PICNIC-defined CNV states, highlighting the increase in M as relative CNV increases. Furthermore, approximate linearity is observed for all qDNA-seq datasets (**Figure 2**), which supports the assumption made by ABCD-DNA in conveying such offsets to the GLM model.

Normalization to "neutral" regions

qDNA-seq read density at any given locus is affected by biological factors, such as CNV and technical factors, such as total sequencing depth and library diversity. Therefore, "normalization" is a subtle yet important aspect for allowing accurate comparison of samples. When are read densities comparable, up to a scaling factor? This question has been addressed in the context of RNA-seq data, where not only expression *level*, but composition of the library and GC content affects read density

(Robinson and Oshlack 2010; Hansen et al. 2012). One popular solution is to use a scaling factor (i.e. an offset) called trimmed mean of M-values (TMM), which allows observations to be kept on their original scale (i.e. counts) for statistical modeling. However, TMM normalization does not explicitly handle CNV or the asymmetry of changes in enrichment (e.g. DNA methylation has opposing global loss in cancer, and localized gain at CpG-rich regions). To estimate normalization factors, we focus on the most prominent "neutral" state. Typically, this will be genomic regions with 2 copies. However, as mentioned, most of the LNCaP genome has 4 copies, so we define neutral as autosomal regions with 2 copies for PrEC and 4 copies for LNCaP (**Figure 1a**); this spans approximately 65% of the reference genome. **Figure 3** shows pairwise comparisons of MBDCap-seq samples using only loci from this neutral state. Due to the logarithm transform, variability of M decreases as A increases (Robinson and Oshlack 2010). However, because of differences in composition and global asymmetry in DNA methylation between samples, the center of the M values does not necessarily occur at 0. Assuming there are regions similarly enriched in both samples, we estimate this bias from "neutral" regions only using the regions of lowest variability (e.g. median of M-values for A > 99^{th} percentile of A-values; See Figure 3) and introduce a sample-specific offset into the statistical model to compensate for expected bias in read densities. Support for this strategy is given in **Supplementary Figure 4**, where normalized data (M-values after adjustment by estimated offsets) for "neutral" loci genome are shown, stratified by CpG density. Despite the asymmetry in DNA methylation, our normalization ensures that the M-value asymptotes are approximately 0, suggesting that read densities are comparable.

Differential calls for various assays and algorithms are positively correlated with CNV

Figures 1B-E highlighted loci where CNV affected read densities, resulting in false or missed detections. To highlight that CNV affects many algorithms genome-wide, we tested several differential approaches: i) DiffBind coupled with MACS output; ii) RSEG; iii) ChIPDiff; iv) ABCD-DNA using 500bp tiled genomic bins. We define *relative rate of peak density* (RRPD) as the number of regions detected in LNCaP divided by the number detected in PrEC, for each CNV state (**Figure 4**). Generally, higher (lower) relative CNV results in more (less) differential region detections, for all algorithms except ABCD-DNA; this positive correlation is indicative of CNV alone affecting the differential calls. Although we do not expect this curve to be completely flat (e.g. interactions between CNV and epigenetics), ABCD-DNA largely removes this association.

Furthermore, CNV may impact many cancer datasets and algorithms. For example, an independent comparison of the LNCaP and PrEC methylome (Kim et al. 2011) by running a region detection algorithm and simply overlapping lists is strongly affected by CNV (**Supplementary Figure 5**). Similarly, differentially methylated regions detected by MeDIP-seq in breast cancer cell lines (Ruike et al. 2010) are associated with CNV, according to their input samples (**Supplementary Figure 6**). Taken together, these results suggest that a non-trivial fraction of differential peak detections could be driven simply by CNV, not changes in relative biological enrichment.

CNV offsets improve differential detection performance

To illustrate that the CNV and normalization offsets proposed above can improve differential detection, we use an independent readout of differential methylation on the same LNCaP and PrEC cells. Using Illumina HumanMethylation 450k BeadChip

arrays, DNA methylation estimates at individual CpG sites are summarized as beta values (See Methods). For comparison with the MBDCap-seq data, beta values are averaged over technical replicates and regions of interest. Here, regions of interest comprise non-overlapping 500bp tiled genomic segments where 450k probes exist. The averaged beta values are used to label regions as differentially methylated (change in beta > 0.4), not differentially methylated (change in beta < 0.1) or indeterminate (0.1-0.4). GLMs are fitted using the edgeR package with and without CNV offsets (both use normalization offsets) and ranking of regions is according to likelihood ratio test P-values. Other cutoffs for difference in beta values were tested (data not shown) and the results presented here are representative.

Figure 5 shows ROC curves for *symmetrically-chosen* truly differentially methylated regions (See Methods), stratified by copy number state, comparing CNV-aware ("ABCD-DNA", using either SNP arrays or genomic sequencing for CNV offsets) and CNV-unaware GLM strategies ("Naïve"), RSEG and DiffBind (with and without input subtraction) are also compared (See Methods). Taken together, these results highlight several features of our new method: i) gains in performance can be achieved for non-"neutral" regions; ii) the magnitude of performance gain increases as CNV increases; iii) ABCD-DNA performs equally well, regardless of the source of CNV information (Affymetrix SNP 6.0, low coverage genomic sequencing); iv) ABCD-DNA outperforms competing methods.

To understand the difference that CNV compensation makes genome-wide to differential detection calls, **Supplementary Figure 6** gives Venn diagrams showing the overlap of CNV-Aware and Naïve calls (adjusted P-value < .01) by CNV state; as expected, differential calls in the "neutral" regions are unaffected, while the overlap degrades significantly as CNV increases. Furthermore, to highlight how ABCD-DNA

removes the association between differential detection and CNV, **Supplementary Figure 7** shows differential detection Z-scores with and without CNV adjustment, stratified by CNV and by "true" 450k differential status used in the ROC comparisons. Naïve scores increase predictably with CNV, whereas ABCD-DNA scores are stable across all CNV states, allowing a better separation of truly differentially methylated from non-differentially methylated.

Because of the asymmetry in the DNA methylation, ROC comparisons are sensitive to the CNV adjustments made. Probes on the 450k arrays are biased towards CpG-rich regions and since these regions often gain methylation in cancer, there is a performance advantage to always increasing the log-fold-change, which can confound the interpretation of the CNV compensation. To eliminate this bias, our results above (**Figure 6**) used randomly selected truly differentially methylated regions such that the same number increased and decreased. However, **Supplementary Figure 8** highlights ROC comparison where this symmetry was not ensured; in this situation, we overstate (understate) performance for lower (higher) relative CNV, as expected.

ABCD-DNA outperforms CNV-aware BATMAN

Next, we compared ABCD-DNA against the CNV-aware BATMAN for the differential analysis of MeDIP-seq data. In the original analysis, read densities were first pre-processed (divided by CNV, explicitly assuming a direct unit slope relationship) to adjust for CNV before using BATMAN (Feber et al. 2011). Their dataset comprises MeDIP-seq, Affymetrix SNP 6.0 and Illumina HumanMethylation 27k arrays for three pooled populations: i) *cancer* versus *normal* (malignant peripheral nerve sheath tumors versus normal Schwann cells); ii) *benign* versus *normal* (benign neurofibromas versus Schwann cells); and, *cancer* versus *benign*. We

use the 27k array data as independent "truth" for our performance evaluation (as above, change in beta > 0.4 defines differentially methylated and change in beta < 0.1 is deemed non-differentially methylated). We estimated CNV offsets from their Affymetrix SNP 6.0 data using PICNIC and normalization offsets using CNV-neutral regions, as above. Notably, because these are sample mixtures, the CNV estimates could be non-integer-valued. **Figure 6** shows ROC curves for the 3 comparisons using 3 differential detection approaches: i) the CNV-aware BATMAN (Down et al. 2008; Feber et al. 2011) ("BATMAN"); ii) count-based analysis with only normalization offsets ("Naïve"); and, iii) count-based analysis with normalization and CNV offsets ("ABCD-DNA"). Overall, these results suggest that two gains in performance can be made: i) count-based methods outperform CNV-aware BATMAN on 2 out of 3 comparisons, perhaps suggesting that modeling the data on its count scale followed by direct comparison of read densities performs well; ii) directly integrating CNV information gives a performance advantage. In addition, BATMAN is specific to methylated DNA capture assays, whereas ABCD-DNA can be applied to other qDNA-seq assays.

DISCUSSION

CNV affects read densities for various qDNA-seq assays. For differential comparisons between cancer and normal epigenomes, results can be both driven and masked by CNV, thus leading to false positives and reduced power (**Figure 1**). Cancer qDNA-seq datasets are on the rise and many will ultimately be affected by CNV. We present a straightforward solution that explicitly models CNV in a well-established count-based framework. Our method, called ABCD-DNA, estimates CNV and normalization offsets, and includes them directly in a GLM, similar to recent approaches applied to RNA sequencing data (Hansen et al. 2012). Thus, we

enable a strategy that jointly accounts for effective sequencing depth and CNV, within statistical models that handle biological replication. We verified the approximately linear relationship between CNV and qDNA-seq on multiple cell line datasets, suggesting that offsets are presented on an appropriate scale to modify the mean response.

Using an independent readout of DNA methylation on 2 datasets, we demonstrated that ABCD-DNA is competitive against existing differential approaches and integrating CNV through offsets can further improve performance. In addition, the ABCD-DNA framework is flexible and extensible. Because a matrix of offsets is matched to the matrix of read densities, there is a facility for analyzing datasets with sample-specific, possibly non-integer, copy number. For example, patient studies, where each has a different copy number profile, could be analyzed. Furthermore, through the offset matrix, the method can adjust for not only CNV and effective sequencing depth, but other technical factors that affect read density, such as GC content or antibody efficiency (Egelhofer et al. 2011; Cheung et al. 2011; Hansen et al. 2012); further study is required to adequately demonstrate this capability for qDNA-seq datasets. Meanwhile, ABCD-DNA can handle replication and complicated experimental designs, since these are already features of the employed model (McCarthy et al. 2012). In principle, ABCD-DNA can make use of any accurate source of CNV information; however, the success of the CNV adjustment is ultimately reliant on the accuracy, resolution and scale of these estimates. Furthermore and perhaps most importantly, ABCD-DNA can be applied to differential analysis of various qDNA-seq datasets, including ChIP-seq.

One potential disadvantage of our approach is the reliance on regions of interest, such as regions tiled along the genome; the positioning of these regions could have some

effect. An alternative strategy would be to consider overlapping bins tiled at high density, in combination with principled techniques for smoothing, such as HMMs, to assemble differential regions; this work is beyond the scope of the proof-of-principle presented here. In addition, ABCD-DNA does not currently have a facility for incorporating "input" or control samples; on our evaluation dataset, DiffBind's explicit input subtraction did not convincingly improve performance and other reports have challenged the appropriateness of such controls (Cheung et al. 2011). Further study is required to make general recommendations on this matter.

The main implication of our results is that CNV information, at least for cancer studies, is required for interpretation of qDNA-seq read densities. Failing to account for CNV may result in false positives and false negatives (e.g. **Figure 1B-E**) and could have significant impact on downstream analyses. For example, if CNV is responsible for a significant fraction of naively determined differentially enriched regions, downstream analyses, such as functional category analysis or pathway analysis, may be confounded by CNV; that is, enriched pathways may largely be a reflection of CNV, not from changes in the epigenetic factor of interest. Since ABCD-DNA adjusts expected read density by number of copies, the method can also facilitate detection of changes in allele-specificity; however, partitioning the reads by allele using genotypes is a more direct approach for this (Statham et al. 2012).

Unfortunately, the requirement for CNV information imposes a potentially costly burden for researchers studying cancer epigenomes, since every sample will need to CNV-typed; this would consume sequencing resources and precious DNA. In practice, the effect of CNV on qDNA-seq can be large or small, depending on the type and severity of the cancers being studied. In the comparison of LNCaP and PrEC cells, the magnitude of CNV change is moderate (most often, changes from 4

copies to 3 or 5), but a large proportion of the genome (~35%) is affected, so significant improvements can be made. Depending on the cancer and the severity, copy number aberrations may be larger in magnitude than our dataset, and affect larger (or smaller) proportions of the genome (Baudis and Cleary 2001). So, the gains to be made from CNV-aware analyses are dataset-dependent. However, from our initial results, there is generally only gains to be made after integrating CNV. Furthermore, while the main motivation to develop ABCD-DNA is to compensate for CNV, we have shown that it performs well relative to existing approaches, so the framework may benefit differential qDNA-seq analyses outside of the cancer field.

METHODS

Estimating CNV from Affymetrix SNP 6.0 microarrays

The PICNIC tool (Greenman et al. 2010), specifically designed for the analysis of Affymetrix SNP 6.0 arrays, was used to estimate absolute copy number genome-wide using default parameters. These regional estimates were matched to the read densities in tiled bins along the genome and used directly as offsets in the downstream CNV-aware GLM count modeling.

Estimating CNV from genomic sequencing

Since read depths in genomic DNA sequencing are affected by local GC content and mappability, we implemented a R routine in the Repitools package (Statham et al. 2010) called `absoluteCN()` that calculates read density, GC content and mappability in bins genome-wide. Bins with mappability less than 75% are removed; a smooth curve is fit to the mode of depth versus GC content. This relationship is removed for each bin by dividing out the fit at the bin's GC content and then scaled

according to knowledge of the most prominent copy state (here, LNCaP=4 and PrEC=2). Read densities are then segmented using CBS (Venkatraman and Olshen 2007).

Choosing regions for ROC analysis "symmetrically"

Because the truly differentially methylated regions for the LNCaP versus PrEC comparison are biased towards hypermethylation, we randomly selected the same number of truly hypermethylated and truly hypomethylated regions for the ROC analysis.

ROC analysis using RSEG

To generate ROC curves for RSEG, we ran `rseg-diff` repeatedly with different values of the `-cdf-cutoff` parameter (between 0.01 and 0.40). For each of the truly differentially methylated and non-differentially methylated regions, the score used for ROC analysis was the maximum cdf-cutoff such that the region was deemed differentially enriched, if at all. See Supplementary website describing the commands used for each tool.

ROC analysis using DiffBind

To generate ROC curves for DiffBind, we set a high P-value threshold when calling `dba.report()`, thus giving scores for the full list of inputted regions. The score used for ranking was the P-value. Furthermore, whether to subtract input reads was controlled by the `bSubControl=FALSE` argument in the call to `dba.analyze()`. Otherwise, default parameters were used.

Processing of Illumina HumanMethylation 450k array data

The HumanMethylation 450k arrays were processed using the R/Bioconductor 'minfi' package using `bg.correct = TRUE` and `normalize = "controls"`, to

generate *beta* values. Differences in beta values were used to determine the truly differentially methylated regions.

Reproducibility of analyses and figures in this manuscript

All data and R code used for the generation of figures in this manuscript are available from <http://imlspenticton.uzh.ch/robinson_lab/ABCD-DNA/> with further description in the Sweave-based Supplementary PDF Document.

DATA ACCESS

Datasets used

The following datasets (with NCBI Gene Expression Omnibus accession numbers) were used for the main comparisons:

1. MBDCap-seq, Affymetrix SNP 6.0 arrays, and low coverage genomic DNA sequencing on LNCaP and PrEC cells and MBDCap-seq of SssI (fully methylated DNA) (GSE24546) (Robinson, Clare Stirzaker, et al. 2010), as well as H3K27me3-seq (GSE38683) and H3K4me3-seq (GSE38682) on the same cell lines.
2. Illumina HumanMethylation 450k bead array on LNCaP and PrEC (GSE34340)
3. From Feber et al. study (Feber et al. 2011), MeDIP-seq, Affymetrix SNP 6.0 arrays and Illumina HumanMethylation 27k were available for pools of malignant peripheral nerve sheath tumors, normal Schwann and benign neurofibromas.

Additional analyses to investigate the association between CNV and differential region detection:

1. Ruike MeDIP-seq and input-seq data (Ruike et al. 2010): reads were downloaded from the DDBJ Sequence Read Archive (accession DRP000030) and remapped to the human hg18 genome. A list of differential regions was obtained from Yoshinao Ruike (personal communication); analysis of the association between their corresponding input-seq read densities and detected differential regions was performed using a custom R script.
2. Kim et al. M-NGS data (Kim et al. 2011): The list of differentially methylated regions was obtained from Mohan Dhanasekaran (personal communication); using our SNP array data (same cell lines), associations were made to their detected regions using a custom R script.

Reproducibility of analyses and figures in this manuscript

All data and R code used for generating figures in this manuscript are available from http://imlspenticton.uzh.ch/robinson_lab/ABCD-DNA/.

Software to run ABCD-DNA

A detailed description of the implementation details for ABCD-DNA is given in the Supplementary PDF Document. Software to run ABCD-DNA is freely available within the Bioconductor Repitools package (Statham et al. 2010).

ACKNOWLEDGEMENTS

We wish to thank Davis McCarthy, Yunshen Chen and Gordon Smyth for early access to the edgeR GLM code and useful discussions; we wish to thank Andrea Riebler and Alicia Oshlack for reading earlier versions of the manuscript, as well as Rory Stark for helpful discussions and Kate Patterson for help with Figures. Funding was from NBCF program (SJC) and NHMRC project grants (CS & SJC).

FIGURE LEGENDS

Figure 1. CNV causes false positives and false negatives to various algorithms; ABCD-DNA can recover them. A) The landscape of CNV between LNCaP (black) and PrEC (grey) cells inferred by PICNIC algorithm (using Affymetrix SNP 6.0 data, see Methods). Using Illumina 450k array data to gauge true differential methylation (See tracks "LNCaP 450k" and "PrEC 450k"), four CNV-induced false positive (FP) or false negative (FN) regions in MBDCap-seq data (See tracks "LNCaP_MBD2" and "PrEC_MBD2") using existing algorithms are shown. Detected differential regions for four methods (ChIPDiff, DiffBind, RSEG, our new approach ABCD-DNA) are shown in black. Region B) shows a FN for all algorithms except ABCD-DNA; the change in depth-normalized read density is not particularly strong, but combined with the knowledge that this is a "low" copy region (LNCaP=2), ABCD-DNA expects fewer reads. Hence, the effective difference is made larger and therefore deemed differential by ABCD-DNA. Similarly, region C) is amplified in cancer beyond "neutral" (LNCaP=5), thus ABCD-DNA expects higher read density (if methylated) and correctly increases the effective change. Region D) is similarly amplified, which causes existing algorithms to overstate the differential methylation (i.e. a FP); note the upstream differentially methylated region that all algorithms detect, whereas only ABCD-DNA correctly attributes the downstream change in read density to CNV. Region E) is lower copy in LNCaP cells, resulting in lower read depth and FPs for all methods except ABCD-DNA.

Figure 2. Linearity between CNV and qDNA-seq. Relative read densities scale linearly with CNV for multiple LNCaP/PrEC qDNA-seq (MBDCap, H3K27me3, H3K4me3) datasets. Scaling factors were calculated separately as the median of log-fold-changes (median of M values) for each CNV stratum and each dataset (See

Supplementary Figure 3); these medians were exponentiated and scaled according to the most prominent CNV state (L=4 P=2). Note that these scaling factors are not actually used in the ABCD-DNA method; they are shown here only to illustrate the relationship between qDNA-seq and CNV.

Figure 3. Normalization to "neutral" CNV state using estimated scaling factors. M (depth-normalized log-fold-change) versus A (depth-normalized average-log) "smear" plots for MBDCap-seq data are shown between technical replicates (A) and between cancer and normal (B); each dot represents a 500bp region of the genome. M is defined as the log-fold-change between 2 samples (counts divided by library size); A is average of the log counts divided by library size. Blue lines represent 99th percentile of A values; red lines denote scale factor estimates (median of M for regions with A greater than 99th percentile). Note: these scale factors are presented here

Figure 4. Association between differential peak detection and CNV across LNCaP/PrEC qDNA-seq datasets using various algorithms. The *relative rate of peak detection* (RRPD), defined as the ratio of the number of regions detected in LNCaP (L) cells to the number of regions detected in PrEC (P), within each CNV stratum is shown for ChIPDiff, RSEG, DiffBind (with and without input subtraction) and ABCD-DNA. DiffBind is based on MACS-detected regions. A) MBDCap-seq; B) H3K27me3-seq; C) H3K4me3-seq. Due to lack of replication, DiffBind was not run on H3K4me3-seq.

Figure 5. ABCD-DNA outperforms competing approaches. ROC curves (sensitivity versus 1-specificity) are shown for various differential region detection algorithms operating on MBDCap-seq data, using 450k array data as an independent source of truly and non-truly differentially methylated regions. "Naïve" uses offsets to account for (effective) sequencing depth but not CNV; "ABCD-DNA" uses either Affymetrix SNP 6.0 or genomic sequencing to estimate CNV offsets. "RSEG" denotes running rseg-diff with different sensitivity cutoffs. "DiffBind", which operates on MACS-detected regions, was run both with and without input subtraction. Each panel shows ROC curves for the respective CNV stratum (between LNCaP and PrEC cells), as indicated in the panel title; the number of such regions is shown in parentheses. In the "L=4 P=2" panel, Naïve and both ABCD-DNA curves almost completely overlap, as do the two DiffBind curves (with and without input subtraction).

Figure 6. ABCD-DNA outperforms CNV-aware BATMAN. ROC curves (sensitivity versus 1-specificity) for 3 pairwise comparisons are shown for a MeDIP-seq dataset (Feber et al. 2011), where Illumina HumanMethylation 27k data is used as an independent source of truly and non-truly differentially methylated regions. "BATMAN" refers to the CNV-adjusted read densities before running the BATMAN algorithm and taking differences in methylation estimates. "Naïve" refers to a count-based analysis, without accounting for CNV. "ABCD-DNA" refers to a count-based analysis, with additional offsets to account for CNV (estimated from Affymetrix SNP 6.0 data using the PICNIC algorithm). Comparisons are: A) cancer versus normal; B) cancer versus benign; C) benign versus normal.

TABLES

Table 1. Table of acronyms for relevant assays and tools.

Acronym	Description	Reference
MBDCap	Methyl-binding domain based capture	-
qDNA-seq	Sequencing of captured DNA subpopulations (i.e. quantitative)	-
GLM	Generalized linear model	(McCarthy et al. 2012)
RSEG	Identifying dispersed epigenomic domains from ChIP-Seq data	(Qiang Song and Smith 2011)
ZINBA	Zero-Inflated Negative Binomial Algorithm	(Rashid et al. 2011)
DiffBind	Differential Binding Analysis of ChIP-Seq peak data	(Ross-Innes et al. 2012)
DBChip	Detecting differential binding of transcription factors with ChIP-seq	(Liang and Keles 2011)
BATMAN	A Bayesian Tool for Methylation Analysis	(Down et al. 2008; Feber et al. 2011)
PICNIC	Predict integral copy numbers in cancer	(Greenman et al. 2010)

REFERENCES

Anders S, and Huber W. 2010. Differential expression analysis for sequence count data. *Genome biology* **11**: R106.

Bardet AF, He Q, Zeitlinger J, and Stark A. 2011. A computational pipeline for comparative ChIP-seq analyses. *Nature Protocols* **7**: 45-61.

Baudis M, and Cleary ML. 2001. Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics* **17**: 1228-1229.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology* **28**: 1045-1048.

Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, Jäger N, Gnirke A, Stunnenberg HG, and Meissner A. 2010. Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature biotechnology* **28**: 1106-14.

Cheung M-S, Down TA, Latorre I, and Ahringer J. 2011. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic acids research* **39**: 1-9.

Curtis C, Lynch AG, Dunning MJ, Spiteri I, Marioni JC, Hadfield J, Chin S-F, Brenton JD, Tavaré S, and Caldas C. 2009. The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics* **10**: 588.

Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM, et al. 2008. A Bayesian deconvolution strategy for

immunoprecipitation-based DNA methylome analysis. *Nature biotechnology* **26**: 779-85.

Egelhofer TA, Minoda A, Klugman S, Lee K, Kolasinska-Zwierz P, Alekseyenko AA, Cheung M-S, Day DS, Gadel S, Gorchakov AA, et al. 2011. An assessment of histone-modification antibody quality. *Nature Structural & Molecular Biology* **18**: 91-93.

Feber A, Wilson GA, Zhang L, Presneau N, Idowu B, Down TA, Rakyan VK, Noon LA, Lloyd AC, Stupka E, et al. 2011. Comparative methylome analysis of benign and malignant peripheral nerve sheath tumors. *Genome research* **21**: 515-24.

Greenman CD, Bignell G, Butler A, Edkins S, Hinton J, Beare D, Swamy S, Santarius T, Chen L, Widaa S, et al. 2010. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics* **11**: 164-75.

Hansen KD, Irizarry RA, and Wu Z. 2012. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 204-216.

Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, Edsall LE, Kuan S, Luu Y, Klugman S, et al. 2010. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell stem cell* **6**: 479-91.

Houseman EA, Christensen BC, Karagas MR, Wrensch MR, Nelson HH, Wiemels JL, Zheng S, Wiencke JK, Kelsey KT, and Marsit CJ. 2009. Copy number variation has little impact on bead-array-based measures of DNA methylation. *Bioinformatics* **25**: 1999-2005.

International Cancer Genome Consortium T. 2010. International network of cancer genome projects. *Nature* **464**: 993-998.

Jones PA, Archer T, Baylin SB, Beck S, Berger S, Bernstein BE, Carpten J, Clark S, Costello J, Doerge R, et al. 2008. Moving AHEAD with an international human epigenome project. *Nature* **454**: 711-715.

Jones PA, and Baylin SB. 2007. The epigenomics of cancer. *Cell* **128**: 683-92.

Kim JH, Dhanasekaran SM, Prensner JR, Cao X, Robinson D, Kalyana-Sundaram S, Huang C, Shankar S, Jing X, Iyer M, et al. 2011. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. *Genome research* **21**: 1028-41.

Liang K, and Keles S. 2011. Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics* 1-2.

McCarthy DJ, Chen Y, and Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research* 1-10.

Rashid N, Giresi PG, Ibrahim JG, Sun W, and Lieb JD. 2011. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biology* **12**: R67.

Robinson MD, McCarthy DJ, and Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140.

Robinson MD, and Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* **11**: R25.

Robinson MD, Statham AL, Speed TP, and Clark SJ. 2010. Protocol matters: which methylome are you actually studying? *Epigenomics* **2**: 587-598.

Robinson MD, Stirzaker Clare, Statham AL, Coolen MW, Song JZ, Nair SS, Strbenac D, Speed TP, and Clark SJ. 2010. Evaluation of affinity-based genome-wide DNA methylation data: effects of CpG density, amplification bias, and copy number variation. *Genome research* **20**: 1719-29.

Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, Brown GD, Gojis O, Ellis IO, Green AR, et al. 2012. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*.

Ruike Y, Imanaka Y, Sato F, Shimizu K, and Tsujimoto G. 2010. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. *BMC genomics* **11**: 137.

Song Q, and Smith AD. 2011. Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics (Oxford, England)* **27**: 870-1.

Statham AL, Robinson MD, Song JZ, Coolen MW, Stirzaker C, and Clark SJ. 2012. Bisulphite-sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. *Genome Research* **13**: 2012.

Statham AL, Strbenac D, Coolen MW, Stirzaker Clare, Clark SJ, and Robinson MD. 2010. Repitools: an R package for the analysis of enrichment-based epigenomic data. *Bioinformatics* **26**: 1662-1663.

Stratton MR. 2011. Exploring the genomes of cancer cells: progress and promise. *Science (New York, N.Y.)* **331**: 1553-8.

Venkatraman ES, and Olshen AB. 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**: 657-663.

Xu H, Wei C-L, Lin F, and Sung W-K. 2008. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics (Oxford, England)* **24**: 2344-9.

Zhang Y, Liu T, Meyer C a, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**: R137.

Zhou Y-H, Xia K, and Wright F a. 2011. A Powerful and Flexible Approach to the Analysis of RNA Sequence Count Data. *Bioinformatics* **27**: 2672-2678.

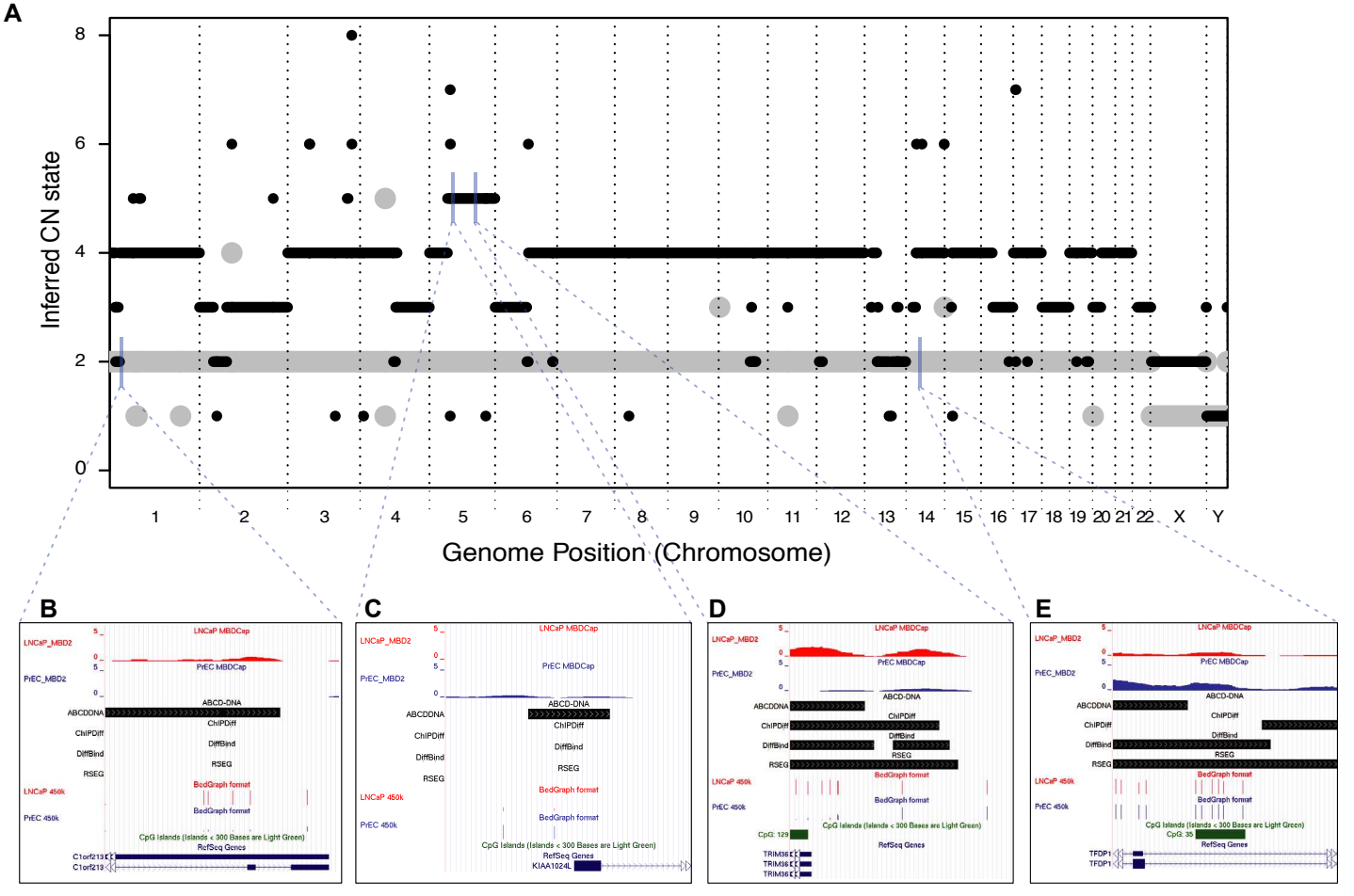
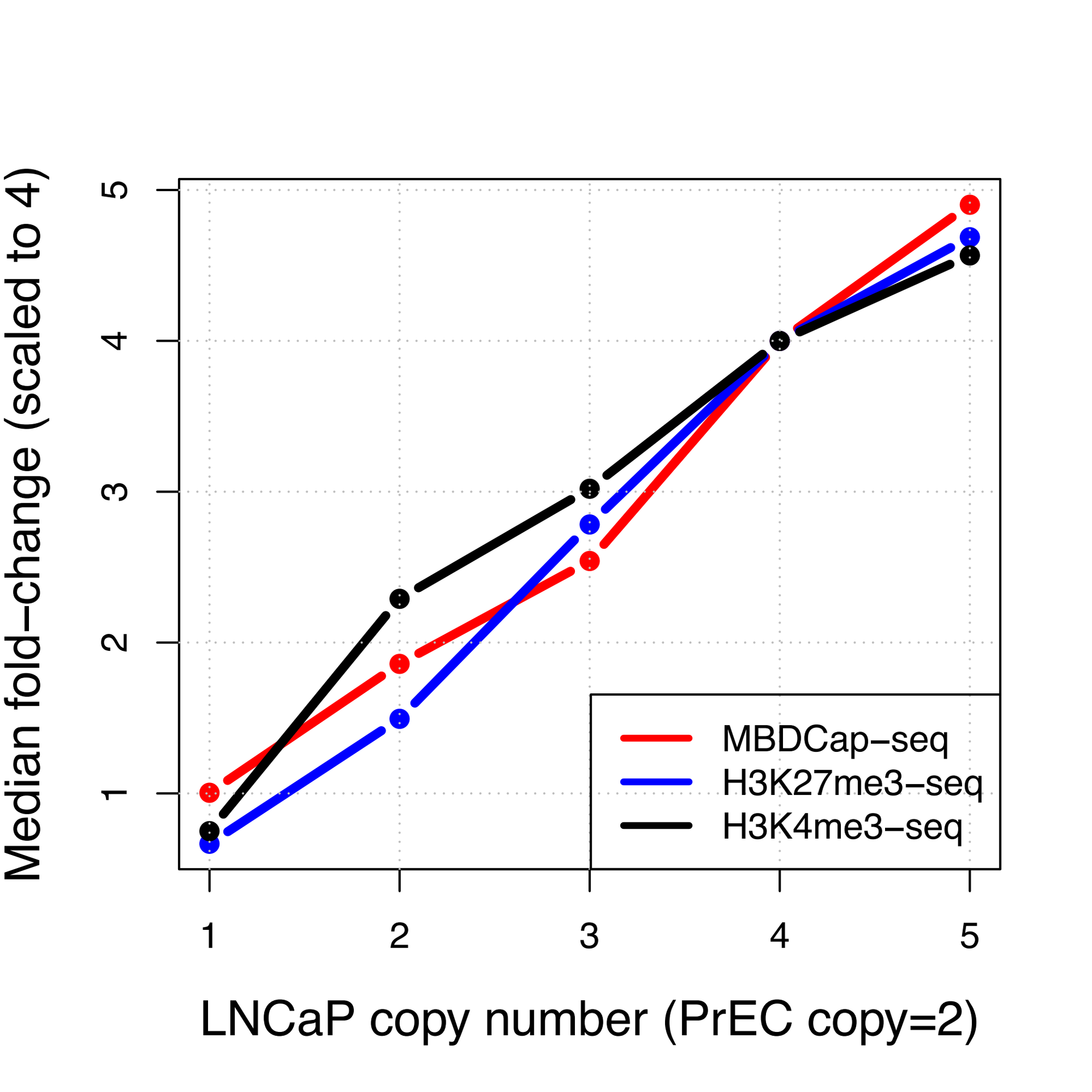
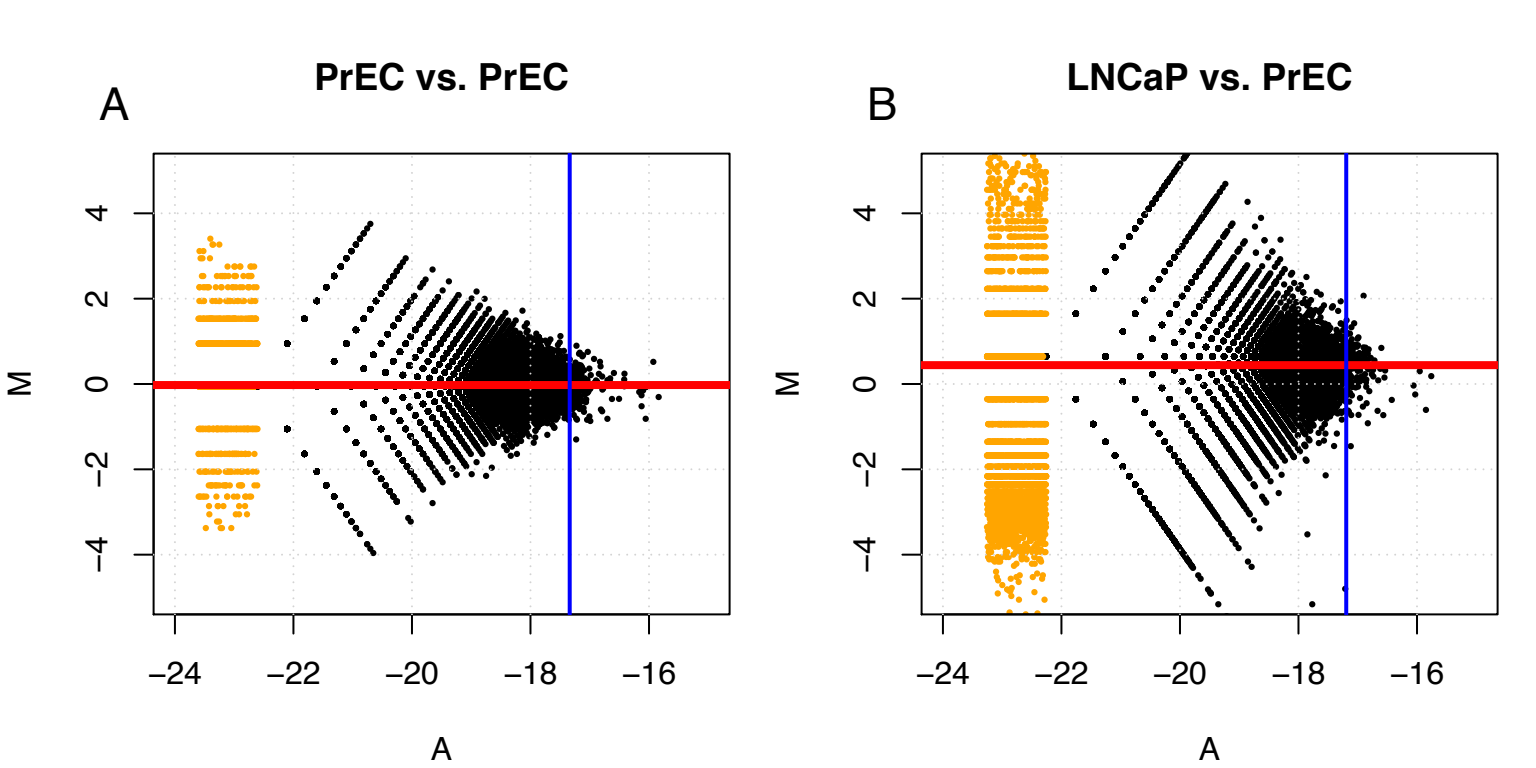
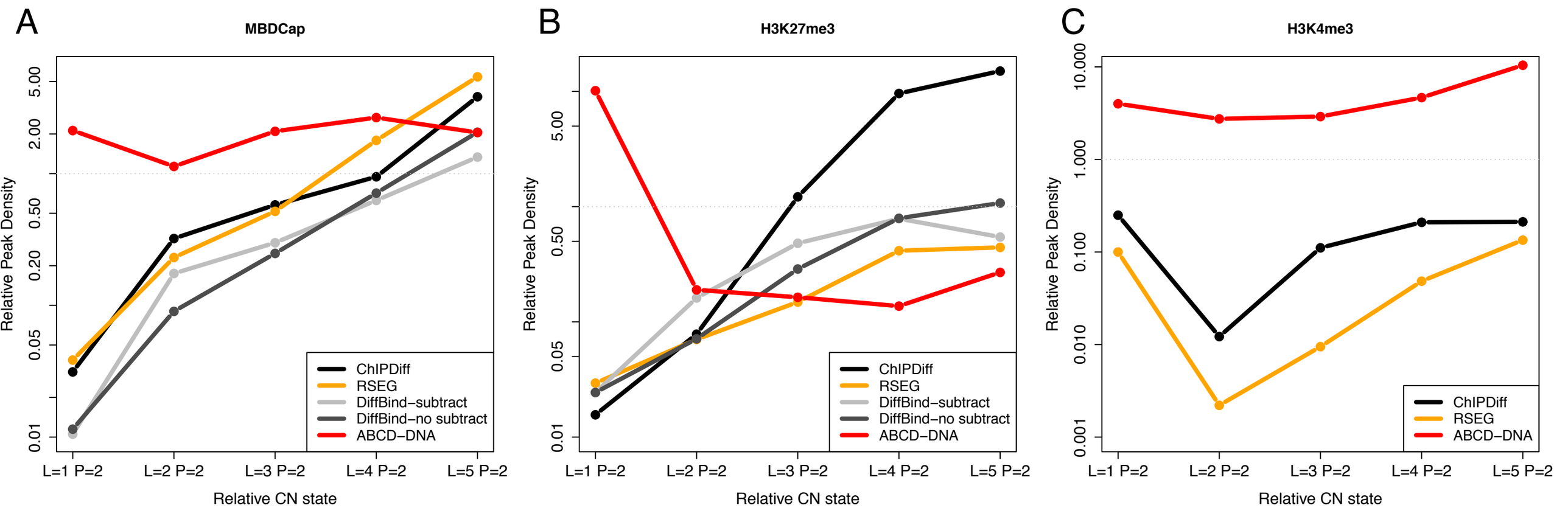
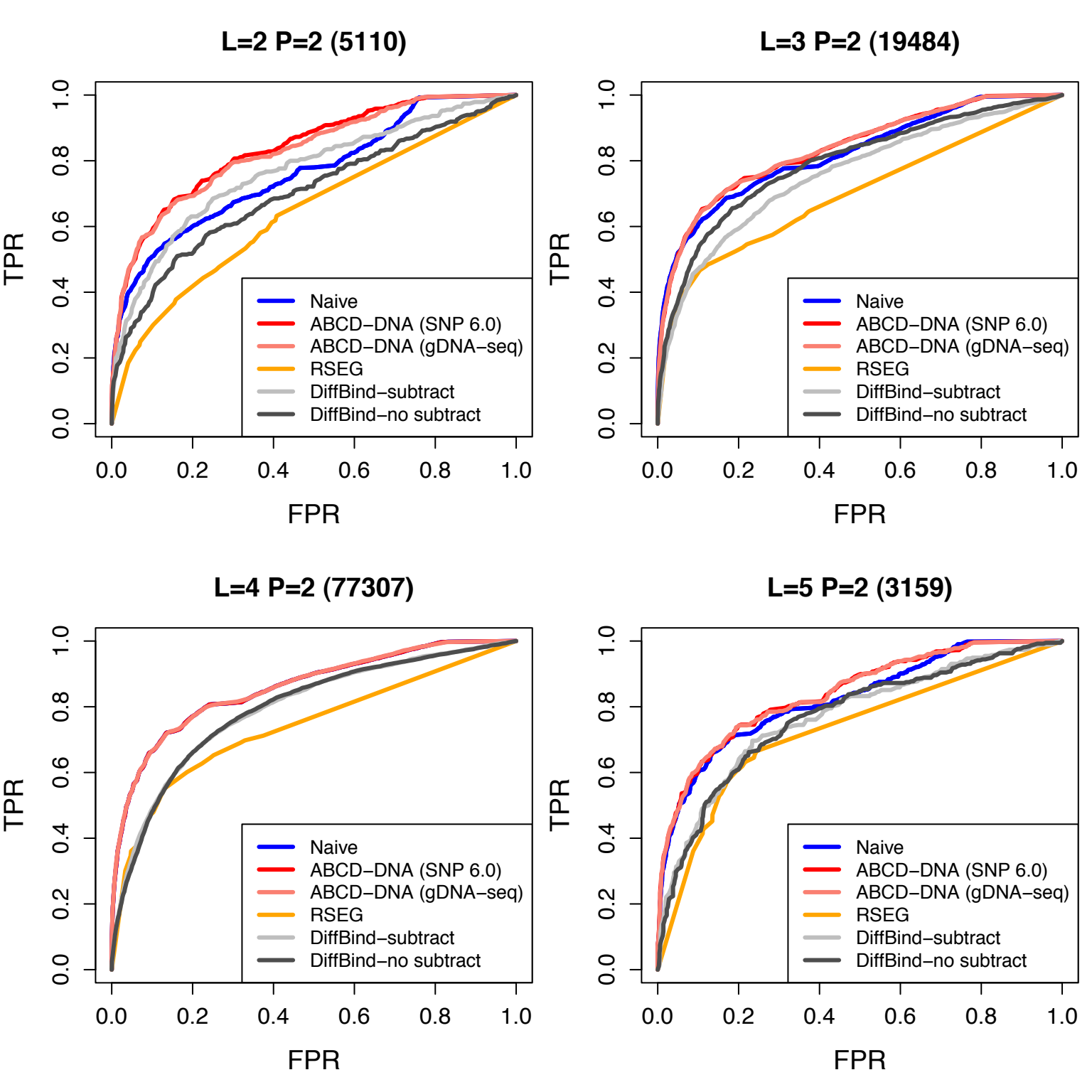

A
Inferred CN state
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Genome Position (Chromosome)
B
C
D
E
LNCaP_MBD2
PrEC_MBD2
LNCaP MBDCap
PrEC MBDCap
ABCD-DNA
ABCDDNA
ChIPDiff
DiffBind
RSEG
BedGraph format
LNCaP 450k
PrEC 450k
CpG Islands (Islands < 300 Bases are Light Green)
RefSeq Genes
C1orf213
KIAA1024L
CpG: 129
TRIM36
CpG: 35
TFDP1

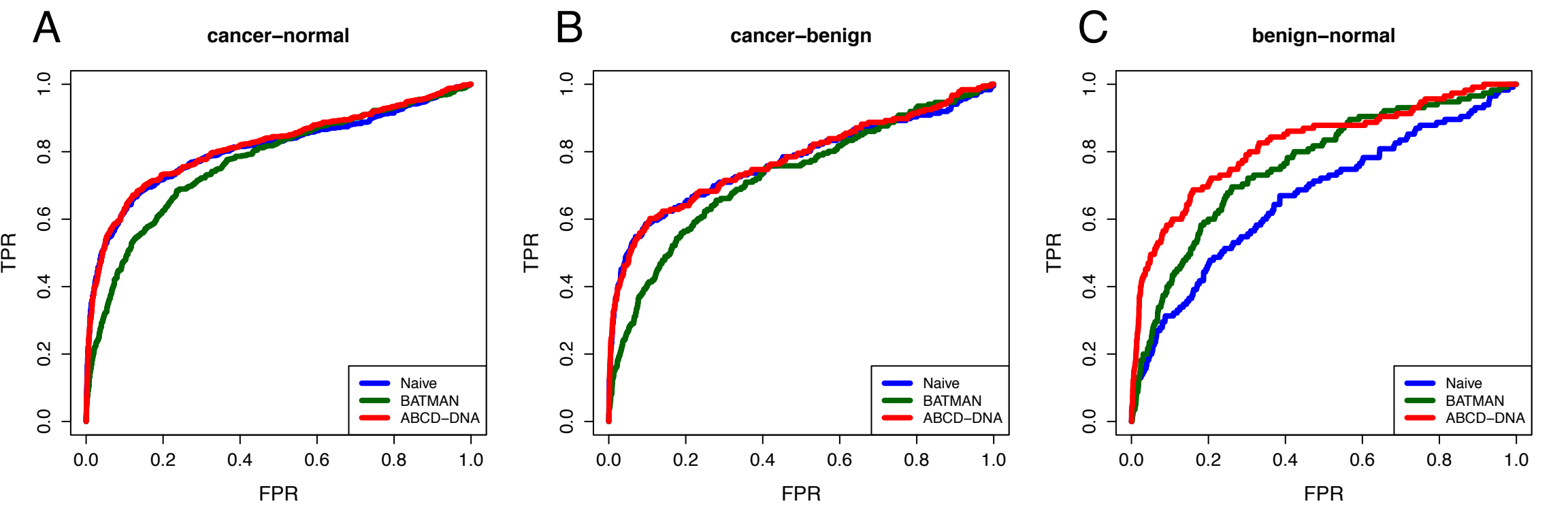
Figure 1


Median fold−change (scaled to 4)
1
2
3
4
5
LNCaP copy number (PrEC copy=2)
1
2
3
4
5
MBDCap−seq
H3K27me3−seq
H3K4me3−seq


A
PrEC vs. PrEC
M
4
2
0
−2
−4
−24
−22
−20
−18
−16
A
B
LNCaP vs. PrEC
M
4
2
0
−2
−4
−24
−22
−20
−18
−16
A


A
MBDCap
Relative Peak Density
0.01
0.05
0.20
0.50
2.00
5.00
L=1 P=2
L=2 P=2
L=3 P=2
L=4 P=2
L=5 P=2
Relative CN state
ChIPDiff
RSEG
DiffBind−subtract
DiffBind−no subtract
ABCD−DNA
B
H3K27me3
Relative Peak Density
0.01
0.05
0.50
5.00
L=1 P=2
L=2 P=2
L=3 P=2
L=4 P=2
L=5 P=2
Relative CN state
ChIPDiff
RSEG
DiffBind−subtract
DiffBind−no subtract
ABCD−DNA
C
H3K4me3
Relative Peak Density
0.001
0.010
0.100
1.000
10.000
L=1 P=2
L=2 P=2
L=3 P=2
L=4 P=2
L=5 P=2
Relative CN state
ChIPDiff
RSEG
ABCD−DNA


L=2 P=2 (5110)
L=3 P=2 (19484)
L=4 P=2 (77307)
L=5 P=2 (3159)
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
Naive
ABCD−DNA (SNP 6.0)
ABCD−DNA (gDNA−seq)
RSEG
DiffBind−subtract
DiffBind−no subtract


A
cancer−normal
B
cancer−benign
C
benign−normal
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
Naive
BATMAN
ABCD−DNA